A simple variable selection technique for nonlinear models

Gianluigi Rech,* Timo Teräsvirta* and Rolf Tschernig**

*Department of Economic Statistics,
Stockholm School of Economics,
Box 6501, S-113 83, Stockholm, Sweden

**Institut für Statistik und Ökonometrie,
Humboldt-Universität zu Berlin,
Spandauer Str. 1, D-10178 Berlin, Germany

February 1, 1999

Abstract

Applying nonparametric variable selection criteria in nonlinear regression models generally requires a substantial computational effort if the data set is large. In this paper we present a selection technique that is computationally much less demanding and performs well in comparison with methods currently available. It is based on a Taylor expansion of the nonlinear model around a given point in the sample space. Performing the selection only requires repeated least squares estimation of models that are linear in parameters. The main limitation of the method is that the number of variables among which to select cannot be very large if the sample is small and an adequate Taylor expansion is of high order. Large samples can be handled without problems.

Keywords: autoregression, nonlinear regression, nonlinear time series, nonparametric variable selection, time series modelling

AMS Classification Code: 62F07

Acknowledgments: The work of the first author has been financed by the Tore Browaldh's Foundation. The second author acknowledges support from the Swedish Council for Research in Humanities and Social Sciences. The research of the third author has been supported by the Deutsche Forschungsgemeinschaft via Sonderforschungsbereich 373 "Quantifikation und Simulation Ökonomischer Prozesse" at Humboldt-Universität zu Berlin.

1. Introduction

Selecting a subset of variables is a problem that has been extensively considered in linear models. The problem often occurs in connection with autoregressive models. In that case, the variables have an ordering, and sequential tests may be applied to choosing the maximum lag if it is finite. If one is also interested in finding the relevant lags, model selection criteria such as FPE (Akaike (1969)), AIC (Akaike (1974)), SBIC (Rissanen (1978); Schwarz (1978)) and many others may be applied.

The variable selection problem also occurs in nonlinear models. In some situations, the functional form of the model may be unknown. The problem of finding the right subset of variables if it exists is then very important. This is because selecting too small a subset leads to misspecification whereas choosing too many variables aggravates the "curse of dimensionality". One way of solving the problem has been to use nonparametric methods based on local estimators. For kernel estimators, Vieu (1995) and Yao and Tong (1994) considered variable selection based on cross-validation. On the other hand, Auestad and Tjøstheim (1990) and Tjøstheim and Auestad (1994) suggested nonparametric FPE criteria. Their technique was further refined by Tschernig and Yang (1998) who, among other things, also showed the consistency of the FPE.

Even when the nonlinear model is parametric, a nonparametric variable selection technique could still be useful in many situations. For instance, if the researcher intends to fit a neural network model to the data, then reducing the dimension of the observation vector before actually fitting any model to data is advisable if possible. Nonparametric variable selection would save the researcher from the effort of estimating a possibly large number of neural network models with different combinations of variables before choosing the final one.

In this paper we propose a simple variable selection procedure that instead of local estimation uses global parametric least squares estimation. It can nevertheless be viewed as a nonparametric procedure as the number of parameters in the global regression is assumed to grow with the number of observations. This approach saves plenty of computational resources compared to nonparametric techniques based on local estimation. It can therefore be easily used even when the number of observations in the time series is large. The plan of the paper is as follows: Section 2 presents the idea and gives the theoretical motivation. Section 3 outlines the model selection procedure. Section 4 reports results from a small-sample simulation study, and Section 5 concludes.

2. Asymptotic motivation

Consider the nonlinear model

$$y_t = f(\mathbf{u}_t; \boldsymbol{\theta}) + \varepsilon_t, t = 1, ..., T \tag{2.1}$$

where $\mathbf{u}_t = (x_{t1}, ..., x_{tp}; z_{t1}, ..., z_{tq})' = (\mathbf{x}_t', \mathbf{z}_t')'$ such that $E[\varepsilon_t \mid \mathcal{F}_t] = 0$, where $\mathcal{F}_t = \{\mathbf{u}_t, \mathbf{u}_{t-1}, ...\}$ is the information set available at time t. Furthermore, we assume that $f(\mathbf{u}_t; \boldsymbol{\theta})$, $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, is a function of $\mathbf{u}_t$ such that it is at least k times continuously differentiable everywhere in the sample space $\mathcal{U} = \{\mathbf{u}_t \mid \mathbf{u}_t \in \mathcal{U}\}$ for all values of $\theta \in \boldsymbol{\Theta}$. Our problem is to find the correct variables (elements of $\mathbf{u}_t$) for the model. Assume that those are $x_{t1}, ..., x_{tp}$, $p \geq 1$, whereas the remaining elements of $\mathbf{u}_t$ are redundant.

We assume that the functional form of f is unknown even if the true function may be parametric. To find the relevant variables, we start by linearizing $f(\mathbf{u}_t; \boldsymbol{\theta})$. This is done by expanding the function into a Taylor series around an arbitrary point $\mathbf{u}_t^0 \in \mathcal{U}$. After merging terms, the k-th order Taylor expansion can be written as:

$$f(\mathbf{x}_t, \mathbf{z}_t; \theta) = \beta_0 + \sum_{j=1}^{p} \beta_j x_{tj} + \sum_{j=1}^{q} \gamma_j z_{tj} + \sum_{j_1=1}^{p} \sum_{j_2=j_1}^{p} \beta_{j_1 j_2} x_{tj_1} x_{tj_2} + \sum_{j_1=1}^{q} \sum_{j_2=j_1}^{q} \gamma_{j_1 j_2} z_{tj_1} z_{tj_2}$$
$$+ \sum_{j_1=1}^{p} \sum_{j_2=1}^{q} \delta_{j_1 j_2} x_{tj_1} z_{tj_2} + \sum_{j_1=1}^{p} \sum_{j_2=j_1}^{p} \sum_{j_3=j_2}^{p} \beta_{j_1 j_2 j_3} x_{tj_1} x_{tj_2} x_{tj_3} + ... +$$
$$+ \sum_{j_1=1}^{q} \sum_{j_2=j_1}^{q} \cdots \sum_{j_k=j_{k-1}}^{q} \gamma_{j_1 ... j_k} z_{tj_1} z_{tj_2} \cdots z_{tj_{k-1}} z_{tj_k} + R_k(\mathbf{u}_t) \tag{2.2}$$

where $q \leq k$ (for notational reasons; this is not a restriction), $R_k(\mathbf{u}_t)$ is the remainder, and the β's, γ's, and δ's are parameters. Expansion (2.2) contains all possible combinations of x_{ti} and z_{ti} up to order k. The assumption that the true data-generating process is only a function of $\mathbf{x}_t = (x_{t1}, ..., x_{tp})'$ means that all the terms involving functions of elements of $\mathbf{z}_t$ in (2.2) have zero coefficients. The remainder term will be a function of $\mathbf{x}_t$ only: $R_k(\mathbf{u}_t) \equiv R_k(\mathbf{x}_t)$, $\forall\ t$. Thus the "true" kth-order expansion is

$$f(\mathbf{x}_t; \boldsymbol{\theta}) = \beta_0 + \sum_{j=1}^{p} \beta_j x_{tj} + \sum_{j_1=1}^{p} \sum_{j_2=j_1}^{p} \beta_{j_1 j_2} x_{tj_1} x_{tj_2} + \sum_{j_1=1}^{p} \sum_{j_2=j_1}^{p} \sum_{j_3=j_2}^{p} \beta_{j_1 j_2 j_3} x_{tj_1} x_{tj_2} x_{tj_3}$$
$$+ ... + R_k(\mathbf{x}_t). \tag{2.3}$$

Equation (2.2) may be written in matrix notation as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}_{\mathbf{x}}\boldsymbol{\gamma} + \mathbf{R}_k(\mathbf{X}, \mathbf{Z}_{\mathbf{x}}) + \boldsymbol{\varepsilon} \tag{2.4}$$

where $\mathbf{X}$ is a $T \times m(k)$ matrix whose $t-$th row involves products of elements of $\mathbf{x}_t$ only, $t = 1, ..., T$, and $\mathbf{Z}_{\mathbf{x}}$ is a $T \times n(k)$ matrix whose $t-$th row consists of elements involving at least one element of $\mathbf{z}_t, t = 1, ..., T$. Setting $\mathbf{W} = \begin{bmatrix} \mathbf{X} & \mathbf{Z}_{\mathbf{x}} \end{bmatrix}$ and $\boldsymbol{\delta} = (\boldsymbol{\beta}', \boldsymbol{\gamma}')'$ and rewriting (2.3) yields

$$\mathbf{y} = \mathbf{W}\boldsymbol{\delta} + \boldsymbol{\varepsilon}^* \tag{2.5}$$

where $\boldsymbol{\varepsilon}^* = \mathbf{R}_k(\mathbf{X}) + \boldsymbol{\varepsilon}$ and $\mathbf{W}$ is of full column rank.

We shall now make additional assumptions about (2.5).Assume that (White (1984), p. 119)

(i) $(\mathbf{w}_t, \varepsilon_t')'$ is a stationary ergodic sequence;

(ii) **(a)** $E\{\mathbf{w}_0\varepsilon_0 \mid \mathcal{F}_{-m}\} \to 0$ in quadratic mean as $m \to \infty$ where $\mathcal{F}_t$ is the information set containing all information about $\mathbf{w}_t$ and $\boldsymbol{\varepsilon}_t$ up until t,

(b) $E \mid \varepsilon_t w_{ti} \mid^2 < \infty$, $i = 1, ..., m(k) + n(k)$,

(c) $\mathbf{V}_T = var\left(T^{-1/2}\mathbf{W}'\boldsymbol{\varepsilon}\right)$ is uniformly positive definite,

(d) $\sum_{j=0}^{\infty} var\left(R_{oij}\right)^{1/2} < \infty, i = 1, ..., m(k) + n(k)$, where $R_{oij} = E\left(w_{0i}\varepsilon_0 \mid \mathcal{F}_{-j}\right) - E\left(w_{0i}\varepsilon_0 \mid \mathcal{F}_{-(j+1)}\right), i = 1, ..., m(k) + n(k)$

(iii) **(a)** $E \mid w_{ti} \mid^2 < \infty$, $i = 1, ..., m(k) + n(k)$,

(b) $\mathbf{M} = E\mathbf{w}_t\mathbf{w}_t'$ is positive definite.

Consider the case where $k, T \to \infty$ such that $(m(k) + n(k))/T \to 0$ as $k \to \infty$. Furthermore, $R_k(\mathbf{x}_t) \to 0$ as $k \to \infty$, for any $k \geq 1$. The OLS estimator of $\boldsymbol{\delta}$ in (2.5) has the form

$$\widehat{\boldsymbol{\delta}} = \left(\mathbf{W}'\mathbf{W}\right)^{-1}\mathbf{W}'\mathbf{y} = \boldsymbol{\delta} + \left(\mathbf{W}'\mathbf{W}\right)^{-1}\mathbf{W}'(\mathbf{R}_k(\mathbf{X}) + \boldsymbol{\varepsilon}) \tag{2.6}$$

where $\boldsymbol{\delta} = \left(\boldsymbol{\beta}', \mathbf{0}'\right)'$. Note that the Taylor approximation becomes arbitrarily accurate as $k \to \infty$; thus $R_{kt} \to 0$ in probability for every t while $(m(k)+n(k))/T \to 0$. As $k \to \infty$, we have

$$\begin{aligned}
\operatorname{p\,lim}_{\substack{T\to\infty \\ k\to\infty \\ (m(k)+n(k))/T\to 0}} (\widehat{\boldsymbol{\delta}} - \boldsymbol{\delta}) &= \operatorname{p\,lim}_{\substack{T\to\infty \\ k\to,\infty \\ (m(k)+n(k))/T\to 0}} \left(\begin{bmatrix} \widehat{\boldsymbol{\beta}} \\ \widehat{\boldsymbol{\gamma}} \end{bmatrix} - \begin{bmatrix} \boldsymbol{\beta} \\ \mathbf{0}_{n(k)} \end{bmatrix} \right) \\
&= \operatorname{p\,lim}_{\substack{T\to\infty \\ k\to\infty \\ (m(k)+n(k))/T\to 0}} \begin{bmatrix} \mathbf{X}'\mathbf{X}/T & \mathbf{X}'\mathbf{Z}_{\mathbf{x}}/T \\ \mathbf{Z}_{\mathbf{x}}'\mathbf{X}/T & \mathbf{Z}_{\mathbf{x}}'\mathbf{Z}_{\mathbf{x}}/T \end{bmatrix}^{-1} \\
&\times \operatorname{p\,lim}_{\substack{T\to\infty \\ k\to\infty \\ (m(k)+n(k))/T\to 0}} \left(\begin{bmatrix} \mathbf{X}'\mathbf{R}_k(\mathbf{X})/T \\ \mathbf{Z}_{\mathbf{x}}'\mathbf{R}_k(\mathbf{X})/T \end{bmatrix} + \begin{bmatrix} \mathbf{X}'\boldsymbol{\varepsilon}/T \\ \mathbf{Z}_{\mathbf{x}}'\boldsymbol{\varepsilon}/T \end{bmatrix} \right) \\
&= \mathbf{M}_{\infty}^{-1} \cdot \operatorname{p\,lim}_{\substack{T\to\infty \\ k\to\infty \\ (m(k)+n(k))/T\to 0}} \begin{bmatrix} \mathbf{X}'\boldsymbol{\varepsilon}/T \\ \mathbf{Z}_{\mathbf{x}}'\boldsymbol{\varepsilon}/T \end{bmatrix} = \mathbf{0}_{\infty}
\end{aligned} \tag{2.7}$$

where the subscript "∞" indicates an infinite-dimensional matrix. Thus, asymptotically, we are able to select the correct set of variables $\mathbf{X}$ with probability

one. Furthermore, Theorem 5.16 (White (1984), p. 119) gives the asymptotic normality of $\sqrt{T}(\hat{\boldsymbol{\delta}} - \boldsymbol{\delta})$. The assumptions we need for these results are rather restrictive in the sense that all moments of $\mathbf{u}_t$ have to exist. However, if $\mathbf{u}_t$ has a multinormal distribution, say, then this assumption is satisfied. We shall discuss the practical implications of our asymptotic theory in Section 3.

The above theory is valid for ordinary regression models, but we would also like to select the appropriate lags in a nonlinear autoregressive model. We can expect our ideas to work in that framework only if we tighten the assumptions about the error structure of the model. Assume that the data-generating process is a nonlinear autoregressive model 2.1 where $\mathbf{u}_t = (y_{t-1}, ..., y_{t-p})'$. It is not sufficient to require that is y_t is stationary. In addition, we also have to assume that at least $2k$ moments of y_t exist if we want to use the kth order Taylor expansion. For asymptotic results similar to those in the previous section, we have to assume that all moments exist. This is the case if $\{\varepsilon_t\}$ is a sequence of zero-mean independent identically distributed stochastic variables such that all moments of their distribution are finite. This can be seen from the Volterra expansion of the autoregressive process; for a definition see Priestley (1981), pp. 869-871.

3. The model selection procedure

The results of Section 2 show that, asymptotically, the combinations containing redundant variables will be discovered as their coefficients that equal zero in the Taylor expansion are estimated consistently as are the other (nonzero) coefficients. The same may not be true for the univariate case, but we have argued that even the factors involving correct variables (lags) contribute more to the Taylor expansion than the other factors. This forms the starting-point of our model selection strategy. It can be described as follows. For a given sample size T, choose k, the order of the Taylor expansion. The asymptotic results suggest that the choice of k is important; k has to be in the right proportion with respect to T. Then regress y_t on all variables (product of original variables) in the Taylor expansion and compute the value of an appropriate model selection criterion. We use SBIC which is a relatively parsimonious criterion that Rissanen (1978) and Schwarz (1978) independently proposed: see, for example, Judge, Griffiths, Hill, Lütkepohl, and Lee (1984), pp. 862-874, or Teräsvirta and Mellin (1986) for other alternatives. Next omit one regressor from the original model, regress y_t on all products of variables remaining in the Taylor expansion and compute the value of SBIC. Repeat this by omitting each regressor in turn. Continue by simultaneously omitting two regressors from the original model. Proceed until the regression only consists of a Taylor expansion of a function with a single regressor. Leave this out as well to check for white noise. This amounts to estimating $\sum_{i=1}^{p+q} \binom{p+q}{i} + 1 = 2^{p+q}$ linear models by ordinary least squares. The combination of variables that yields the lowest value of SBIC is selected. If the number of observations is sufficiently

high and k is selected in an appropriate way, then one should be able to select the correct set of regressors with a high probability.

Sometimes the unknown function in (2.1) may be at least approximately linear. Therefore it may be a good idea to begin the variable selection procedure by testing linearity. This can be done by testing the null hypothesis that the coefficients of all the terms of order higher than one equal zero in the Taylor expansion (2.2). Teräsvirta, Lin, and Granger (1993) arrived at this hypothesis when they derived a test of linearity against a single hidden layer feedforward artificial neural network model. If this hypothesis is not rejected, then the model selection simplifies to variable selection in linear regression using subset regressions. This means saving computer time and making the selection procedure more efficient. As before, a suitable model selection criterion such as SBIC may be applied to the problem. As we assume that the number of variables in f is fixed, SBIC asymptotically yields the correct model with probability one, if the true model is linear.

In the next section we shall apply our procedure to demonstrate how its performs. We compare it with the FPE procedure by Tschernig and Yang (1998) that builds on the work by Tjøstheim and Auestad (1994). That procedure was chosen since it is also consistent while requiring weaker moment assumptions, e.g. the function $f(\cdot)$ only needs to be differentiable up to order four. This is achieved by using local estimation techniques. Instead of increasing the order of the Taylor expansion with increasing sample size as for the global estimator (2.2), the order of the Taylor expansion is fixed while the expansion is estimated only locally. One thus estimates the function value $f(\mathbf{u})$ at $\mathbf{u}$ by estimating a first order expansion with observations lying in a neighbourhood of $\mathbf{u}$. Clearly, the smaller the neighbourhood determined by a so-called bandwidth parameter, the smaller the bias but the larger the estimation variance. With increasing sample size, the approximation error is reduced by decreasing the size of the neighbourhood instead of increasing the order k of the Taylor expansion.

The trade-off between bias and variance allows one to derive an asymptotically optimal bandwidth. Using recent results of Tschernig and Yang (1998), it can be estimated by plug-in methods. A corresponding estimate of k is beyond the scope of this paper. The nonparametric CAFPE proposed by Tschernig and Yang and used in the Monte Carlo analysis is given in equation $(A.5)$ in the Appendix. We shall only report results for univariate models. The results for multivariate models are similar to those for univariate ones and are therefore omitted.

4. A simulation study

To find out how the selection procedure functions in practice, we conducted a simulation study. We simulated both nonlinear autoregressive models and models with exogenous regressors. The autoregressive data-generating processes (DGP) are defined as follows:

(i) Artificial Neural Network model with two lags and a single hidden unit (**ANN1**)

$$Y_t = 0.5 + \frac{1}{1+\exp\{-2(Y_{t-1} - 3Y_{t-2} - 0.05)\}} + \varepsilon_t, \varepsilon_t \sim N\left(0, 10^{-2}\right) \quad (4.1)$$

(ii) Nonlinear Additive AR(2) process (**NLAR1**)

$$Y_t = -0.4 \cdot \frac{3 - Y_{t-1}^2}{1 + Y_{t-1}^2} + 0.6 \cdot \frac{3 - (Y_{t-2} - 0.5)^3}{1 + (Y_{t-2} - 0.5)^4} + 0.1\varepsilon_t, \varepsilon_t \sim N(0,1) \quad (4.2)$$

(iii) Nonlinear Additive AR(4) process(**NLAR1_14**)

$$Y_t = -0.4 \cdot \frac{3 - Y_{t-1}^2}{1 + Y_{t-1}^2} + 0.6 \cdot \frac{3 - (Y_{t-4} - 0.5)^3}{1 + (Y_{t-4} - 0.5)^4} + 0.1\varepsilon_t, \varepsilon_t \sim N(0,1) \quad (4.3)$$

(iv) Nonlinear AR(2) process (**NLAR4**)

$$Y_t = 0.9 \cdot \frac{1}{1 + Y_{t-1}^2 + Y_{t-2}^2} - 0.7 + 0.1\varepsilon_t \quad (4.4)$$

$$\varepsilon_t \sim \text{Triangular density, positive for } |\varepsilon_t| < 0.1\sqrt{6}$$

(v) Logistic smooth transition autoregressive process (**LSTAR)**

$$Y_t = 1.8Y_{t-1} - 1.06Y_{t-2} + (0.02 - 0.90Y_{t-1} + 0.795Y_{t-2})$$
$$\times \frac{1}{1+\exp\{-100(Y_{t-1} - 0.02)\}} + \varepsilon_t, \ \varepsilon_t \sim N\left(0, 10^{-2}\right) \quad (4.5)$$

(vi) Periodic: (**SIN1)**

$$Y_t = \frac{\sin(\pi Y_{t-1}) + \sin(4\pi Y_{t-2})}{2} + \varepsilon_t, \ \varepsilon_t \sim N\left(0, 10^{-2}\right) \quad (4.6)$$

Simulating these processes did not indicate that any of them would be explosive. The random numbers were generated by the random number generator in GAUSS, version 3.2. The first 200 observations of each series were discarded. Models (4.2), (4.3) and (4.6) are additive; models (4.1) , (4.4) , (4.5) are not. This seems to make a difference if we apply the nonparametric FPE procedure of Tschernig and Yang (1998). The LSTAR model (4.5) is the same as that in Teräsvirta (1994), p. 211 except for the error variance, which is greater. The periodic model (4.6) may be expected to be a problematic one as it is not well approximated by a combination of low-order polynomials of its variables. We also simulated models with exogenous regressors with the same functional form as the univariate ones, the lags being replaced by normally distributed exogenous regressors generated by a stationary first-order vector autoregression. The results from both cases are rather similar, and we therefore only report those based on univariate models. The remaining results are available from the authors upon request.

We used three sample sizes, $T = 100, 200, 2000$, in this study. To compare our procedure with the nonparametric approach, we also simulated the CAFPE procedure for the two smallest sample sizes. For $T = 2000$, the computing times for that technique turned out to be prohibitive. The idea with the smallest sample size is precisely to see how our procedure works when the available information set is not very large. The other two sample sizes are chosen (i) to show how much things improve compared to the smallest sample size and (ii) to demonstrate that the choice of k is important, that is, the ratio $(m(k) + n(k))/T$ has to approach zero at a "right" rate as T increases.

Table 1 contains the results for $T = 100$. Most realizations from the nonlinear model (4.4) seem linear, and selecting the correct lags is more difficult than in other models. Note, however, that the ranking of the models in this respect may easily be changed by changing the error variance. Results also indicate that the nonparametric approach works better when the DGP is additive. In that case, CAFPE and our Taylor expansion strategy produce similar results. In other cases, our procedure compares quite favourably with the nonparametric one.

Tables 3 and 4 are based on 2000 observations. A comparison between them and Table 2 shows that the choice of k is important. If $k = 3$ as in Table 2 then, for the additive models, the performance of the procedure deteriorates compared to the smaller sample size. If $T = 2000$ but we choose a fifth-order Taylor expansion, results are uniformly at least as good as for $T = 200$. An overall conclusion from this, admittedly limited, simulation experiment, is that the simple Taylor-expansion based model selection strategy works quite well. But this requires finding the right balance between the sample size, the number of variables under consideration, and the order of the Taylor expansion. Note in particular that for the periodic model (4.6), our technique performs adequately only for $T = 2000$ if, at the same time, $k = 5$. In small samples it is solidly beaten by the CAFPE.

The order of the autoregression p is restricted by the procedure because the number of regressors in the auxiliary regression grows exponentially with p and k. We can alleviate the problem as follows. If we disregard the cross terms in the Taylor expansion, we are able to reduce the number of regressors substantially. This means implicitly assuming that the underlying model is additive. The lags selected this way normally encompass the set we would have selected with the complete expansion, if using that had been possible. Repeating the procedure with the complete Taylor expansion for the selected subset may then weed out the remaining redundant variables.

5. Conclusions

We have developed a variable selection technique that can be applied to nonlinear models and provided an asymptotic justification for it. Time series with a couple of hundred observations at most do not allow the set of variables to choose from to be large. On the other hand, the other techniques available for nonlinear model selection share this disadvantage. One of the main advantages of our technique is that it is simple and computationally feasible because it is based on ordinary least squares. It is applicable already in small samples and the computational burden remains tolerable even when the series are long. The standard subset selection procedure for linear models constitutes a special case of our technique. In small samples the performance of our variable selection procedure compares favourably with currently available techniques based on nonparametric methods.

References

Akaike, H. (1969): "Fitting Autoregressive Models for Prediction," *Ann. Inst. Statist. Math.*, 21, 243–47.

——— (1974): "A New Look at the Statistical Model Identification," *IEEE Trans. Autom. Control*, AC-19, 716–23.

Auestad, B., and D. Tjøstheim (1990): "Identification of Nonlinear Time Series: First Order Characterization and Order Determination," *Biometrika*, 77, 669–87.

Judge, G. G., W. E. Griffiths, R. C. Hill, H. Lütkepohl, and T.-C. Lee (1984): *The Theory and Practice of Econometrics*. New York: Wiley.

Priestley, M. (1981): *Spectral Analysis and Time Series*. New York, Academic Press.

Rissanen, J. (1978): "Modeling by Shortest Data Description," *Automatica*, 14, 465–471.

Schwarz, G. (1978): "Estimating the Dimension of a Model," *Ann. Statist.*, 6, 461–64.

Silverman, B. (1986): *Density Estimation for Statistics and Data Analysis.* London: Chapman and Hall.

Teräsvirta, T. (1994): "Specification, Estimation, and Evaluation of Smooth Transition Autoregressive Models," *Journal of the American Statistical Association*, 89, 208–218.

Teräsvirta, T., C.-F. J. Lin, and C. W. J. Granger (1993): "Power of the Neural Network Linearity Test," *Journal of Time Series Analysis*, **14**, 209–220.

Teräsvirta, T., and I. Mellin (1986): "Model Selection Criteria and Model Selection Tests in Regression Models," *Scandinavian Journal of Statistics*, **13**, 159–171.

Tjøstheim, D., and B. Auestad (1994): "Nonparametric Identification of Nonlinear Time Series: Selecting Significant Lags," *Journal of the American Statistical Association*, **89**, 1410–1419.

Tschernig, R., and L. Yang (1998): "Nonparametric Lag Selection for Time Series," Discussion paper, Humboldt University Berlin.

Vieu, P. (1995): "Order Choice in Nonlinear Autoregressive Models," *Statistics*, **26**, 307–328.

White, H. (1984): *Asymptotic Theory for Econometricians.* Orlando, FL: Academic Press.

Yao, Q., and H. Tong (1994): "On Subset Selection in Non-Parametric Stochastic Regression," *Statistica Sinica*, **4**, 51–70.

Appendix

Denote by $\mathbf{w}$ a $(m \times 1)$ subvector of $\mathbf{u}$, $m \leq p+q$. A local linear estimate $\widehat{f}(\mathbf{w}, h)$ of the function at $\mathbf{w}$ using the bandwidth h is given by the estimated constant $\hat{c}_0$ of a linear Taylor expansion fitted locally around $\mathbf{w}$

$$\{\widehat{c}_0, \widehat{\mathbf{c}}\} = \arg\min_{\{c_0, \mathbf{c}\}} \sum_{t=1}^{T} \{y_t - c_0 - (\mathbf{w}_t - \mathbf{w})'\mathbf{c}\}^2 K_h(\mathbf{w}_t - \mathbf{w}) \qquad (A.1)$$

where $K(\cdot)$ denotes a standard kernel function and $K_h(\mathbf{w}_t - \mathbf{w}) = h^{-m} \prod_{i=1}^{m} K((w_{t,i} - w_i)/h)$. The integrated mean squared error can then be estimated by

$$\widehat{A}(h) = T^{-1} \sum_{t=1}^{T} \left\{ y_t - \widehat{f}(\mathbf{w}_t, h) \right\}^2 w(\mathbf{u}_t) \qquad (A.2)$$

where the integration is restricted to the domain of the weight function $w(\cdot)$ which is defined for the full vector $\mathbf{u}$. Furthermore, define the term

$$\widehat{B}(\widehat{h}_B) = T^{-1} \sum_{t=1}^{T} \left\{ y_t - \widehat{f}(\mathbf{w}_t, \widehat{h}_B) \right\}^2 w(\mathbf{u}_t)/\widehat{\mu}(\mathbf{w}_t, \widehat{h}_B) \qquad (A.3)$$

where $\widehat{\mu}(\cdot)$ is a Gaussian kernel estimator of the density μ using Silverman's (Silverman (1986)) rule-of-thumb bandwidth $\widehat{h}_B = h(m+2, \hat{\sigma}, T)$ and

$$h(k, \sigma, n) = \sigma\{4/k\}^{1/(k+2)} n^{-1/(k+2)}.$$

Moreover, $\widehat{\sigma} = \left(\prod_{j=1}^{m} \sqrt{Var(\mathbf{w}_j)} \right)^{1/m}$ denotes the geometric mean of the standard deviation of the regressors.

The local linear estimate of the FPE is then given by

$$AFPE = \widehat{A}(\widehat{h}_{opt}) + 2K(0)^m T^{-1} \widehat{h}_{opt}^{-m} \widehat{B}(\widehat{h}_B) \qquad (A.4)$$

where the plug-in bandwidth is computed from

$$\widehat{h}_{opt} = \left\{ m||K||_2^{2m} \widehat{B}(\widehat{h}_B) T^{-1} \widehat{C}(\widehat{h}_C)^{-1} \sigma_K^{-4} \right\}^{1/(m+4)}$$

with $||K||_2^2 = \int K^2(u)du$, $\sigma_K^2 = \int K(u)u^2 du$. Note that the second term in $(A.4)$ serves as a penalty term to punish overfitting.

The estimation of C involves second derivatives which are estimated with a local quadratic estimator that excludes all cross derivatives. It is a simplification of the partial local cubic estimator of Tschernig and Yang (1998). The bandwidth estimate $\widehat{h}_C$ is given by $h(m+4, 3\widehat{\sigma}, T)$.

Based on theoretical reasons and Monte Carlo evidence provided in Tschernig and Yang (1998), the authors suggest to use the corrected FPE

$$CAFPE = AFPE \left\{ 1 + mT^{-4/(m+4)} \right\} \qquad (A.5)$$

where the correction increases the probability of correct fitting. One then chooses that variable vector $\mathbf{w}^*$ for which $AFPE$ or $CAFPE$ are minimized.

Table 1. The number of correct choices among the first **5** lags (**C**), number of underfitted models (**U**; fewer lags than in the correct model), number of overfitted models (**O**; more lags than in the correct model) in 100 replications with $T = 100$ observations. Order of the Taylor expansion = 3.

Model	Correct lags	Taylor expansion method							CAFPE
		Linear			Nonlinear			Lin + Nonlin	
		C	U	O	C	U	O	C	C
ANN1	1,2	25	3	2	65	5	0	90	24
NLAR1	1,2	0	0	0	98	1	1	98	99
NLAR1_14	1,4	0	0	0	98	0	2	98	99
NLAR4	1,2	61	32	5	0	2	0	61	25
LSTAR1	1,2	57	5	0	31	7	0	88	46
SIN1	1,2	0	51	4	0	45	0	0	55

Table 2. The number of correct choices among the first 6 lags (**C**), number of underfitted models (**U**; fewer lags than in the correct model), number of overfitted models (**O**; more lags than in the correct model) in 100 replications with $T = 200$ observations. Order of the Taylor expansion = 3.

Model	Correct lags	Taylor expansion method							CAFPE
		Linear			Nonlinear			Lin + Nonlin	
		C	U	O	C	U	O	C	C
ANN1	1,2	0	0	0	100	0	0	100	36
NLAR1	1,2	0	0	0	99	0	1	99	99
NLAR1_14	1,4	0	0	0	99	0	1	99	100
NLAR4	1,2	95	1	3	0	1	0	95	52
LSTAR1	1,2	13	0	0	86	1	0	99	43
SIN1	1,2	0	6	0	0	94	0	0	98

Table 3. The number of correct choices among the first 6 lags (**C**), number of underfitted models (**U**; fewer lags than in the correct model), number of overfitted models (**O**; more lags than in the correct model) in 100 replications with $T = 2000$ observations. Order of the Taylor expansion = 3.

Model	Correct lags	Taylor expansion method						
		Linear			Nonlinear			Lin + Nonlin
		C	U	O	C	U	O	C
ANN1	1,2	0	0	0	100	0	0	100
NLAR1	1,2	0	0	0	62	0	38	62
NLAR1_14	1,4	0	0	0	20	0	80	20
NLAR4	1,2	69	0	3	28	0	0	97
LSTAR1	1,2	0	0	0	100	0	0	100
SIN1	1,2	0	0	0	9	91	0	9

Table 4. The number of correct choices among the first 6 lags (**C**), number of underfitted models (**U**; fewer lags than in the correct model), number of overfitted models (**O**; more lags than in the correct model) in 100 replications with $T = 2000$ observations. Order of the Taylor expansion = **5**.

Model	Correct lags	Taylor expansion method						
		Linear			Nonlinear			Lin + Nonlin
		C	U	O	C	U	O	C
ANN1	1,2	0	0	0	100	0	0	100
NLAR1	1,2	0	0	0	100	0	0	100
NLAR1_14	1,4	0	0	0	100	0	0	100
NLAR4	1,2	59	0	1	40	0	0	99
LSTAR1	1,2	0	0	0	100	0	0	100
SIN1	1,2	0	0	0	94	6	0	94